Certified Big Data Science Analyst (CBDSA)

www.globalicttraining.com

DURATION

- 5 Days

COURSE OBJECTIVES

The statistic shows a revenue forecast for the global big data industry from 2011 to 2026. For 2017, the source projects the global big data market size to grow to just under 34 billion U.S. dollars in revenue (https://www.statista.com/statistics/254266/global-big-data-market-forecast/)

The creation and consumption of data continues to grow by leaps and bounds and with it the investment in big data analytics hardware, software, and services and in data scientists and their continuing education. The availability of very large data sets is one of the reasons Deep Learning, a sub-set of artificial intelligence (AI), has recently emerged as the hottest tech trend, with Google, Facebook, Baidu, Amazon, IBM, Intel, and Microsoft, all with very deep pockets, investing in acquiring talent and releasing open AI hardware and software.

This course will transfer Technical know-how about the concept of Business Analytics and its importance in today's market. Participants will acquire knowledge on different Data Mining techniques and tool (RapidMiner). This course objective is to introduce participants about the Big Data Solution (Hadoop) and the components working on top of Hadoop
(HBase, Hive).

By the end of this course, participants will have a good understanding of how the Big Data storage and processing works to accomplish today's growing need to work on all variety and volume of data.

As part of the course, participant will be given a case study and it would cover all the aspects of the Business Analytics and Big Data covered in the course. Participants will be required to give a solution to the problem using all components taught in the course.

JOB ROLES IN NICF / TARGETED AUDIENCE

- Data Analyst - Statistics and Mining
- Big Data Analyst
- Operations Research Analyst
- Data Scientist
- IHL students

PRE-REQUISITES

Participants are recommended to have experience in software development with Java/Unix/ Linux environment and good understanding on data/business analytics.

PROGRAM STRUCTURE

Certified Big Data Science Analyst program is a 5-day intensive training program with the following assessment components.

Component 1. Written Examination
Component 2. Project Work Component (PWC)

These components are individual based. Participants will need to obtain 70% in both the components in order to qualify for this certification. If the participant fails one of the components, they will not pass the course and have to re-take that particular failed component. If they fail both components, they will have to re-take the assessment.

COURSE SESSION SCHEDULE

Day 1	**Session 1 (9:00 – 10:30)**	**Session 2 (10:40 – 12:10)**	**Session 3 (13:10 – 14:10)**	**Session 4 (14:10 – 18:10)**
	Introduction to Business Analytics	Introduction to Business Analytics	Introduction to Business Analytics	Data/Information Architecture for Business Analytics
Day 2	**Session 1 (9:00 – 10:30)**	**Session 2 (10:40 – 12:10)**	**Session 3 (13:10 – 14:10)**	**Session 4 (14:10 – 18:10)**
	Data Mining Tool	Data Mining Tool	Data Mining Tool	Data Mining Techniques
Day 3	**Session 1 (9:00 – 10:30)**	**Session 2 (10:40 – 12:10)**	**Session 3 (13:10 – 14:10)**	**Session 4 (14:10 – 18:10)**
	Introduction to Big Data	Introduction to Big Data	Introduction to Big Data	Introduction to Hadoop
Day 4	**Session 1 (9:00 – 10:30)**	**Session 2 (10:40 – 12:10)**	**Session 3 (13:10 – 13:40)**	**Session 4 (13:40 – 18:40)**
	Hadoop HDFS & MapReduce	Hadoop HDFS & MapReduce	Hadoop HDFS & MapReduce	Apache HBase
Day 5	**Session 1 (9:00 – 10:00)**	**Session 2 (10:10 – 12:10)**	**Session 3 (13:10 – 15:10)**	**Session 4 (15:10 – 17:40)**
	Apache Hive	Apache Hive	Apache Hive	CBDSA examination

COURSE OUTLINE

Unit 1: Introduction to Business Analytics

- The concept of Business Analytics
- Data, Information, Knowledge and Wisdom
- Data as Unique Enterprise Asset
- Data, Information and Analytics Lifecycle
- Business Analytics – Current Context
- Types of Analytics
    - Descriptive Analytics
    - Predictive Analytics
    - Prescriptive Analytics

Unit 2: Data/Information Architecture for Business Analytics

- Data/Information Architecture
- Concept of Data Warehouse/Enterprise Data Warehouse (EDW)
- ETL – Key Process
- Concept of Data Mart
- Business Intelligence
- Data Mining

Unit 3: Data Mining Tool

- Understand the open source DM tool RapidMiner
- Explore the various features of RapidMiner
- Walkthrough a RapidMiner demo with different scenarios

Unit 4: Data Mining Techniques

- Understand the various data mining techniques
- Understand how correlation matrix works
- Understand how association rule mining works
- Understanding the Predictive Analytics technique
- Understand the forecasting technique

Unit 5: Introduction to Big Data

- What is Big Data? Why Big Data?
- 3V's of Big Data
- The Rapid Growth of Unstructured Data
- Big Data Market Forecast
- Big Data Analytics
- Big Data in Business
- Big Data Types & Architecture

Unit 6: Introduction to Hadoop

- Big Data – Current Industry Trends
- Why Process Big Data?
- Challenges in Data Processing
- Why Hadoop?
- What is Hadoop offering?
- Hadoop Network Structure
- Hadoop Eco-System
- Hadoop Core Components
- Hadoop – Features
- Hadoop – Relevance
- Hadoop in Action
- Sqoop import and export

Unit 7: Hadoop HDFS & MapReduce

- Hadoop HDFS
    - What does HDFS Facilitate?
    - HDFS Architecture
    - Hadoop Network and Server Infrastructure
    - NameNode, Secondary NameNode and DataNode
    - Ensuring Data Correctness
    - Data Pipelining while Loading Data
    - fs Operations
- Hadoop MapReduce
    - MapReduce Conceptualization
    - MapReduce – Overview
    - MapReduce – Programming Model
    - MapReduce – Execution Overview
    - Hadoop – Application Examples
    - Word Count – Example

Unit 8: Apache HBase

- What is HBase?
- HBase Architecture
- ZooKeeper
- HBase Data model
- HBase Deployment
- HBase Cluster Architecture
- Indexes in HBase
- Scaling HBase
- Data Locality, Coherence and Concurrency, Fault Tolerance
- Hadoop Integration

- High-Level Architecture
- Replication of Data Across Data Centres
- HBase Applications
- Advantages and Disadvantages

Unit 9: Apache Hive

- What is Hive?
- Why Hive?
- Where to use Hive?
- Hive Architecture
- Hive: Benefits
- Hive: Tradeoffs
- Hive: Real world Examples

WRITTEN ASSESSMENT

As part of the written examination, each participant will be assessed individually on the last day of the training for their understanding of the subject matter and ability to evaluate, choose and apply them in specific context and also the ability to identify and manage risks. The assessment focuses on higher levels of learning in Bloom's taxonomy: Application, Analysis, Synthesis and Evaluation.

This written examination will primarily consist of 40 multiple choice questions spanning various aspects as covered in the program. It is an individual, competency-based assessment.

COURSE OUTCOME

- Acquire knowledge of complete Big Data Technologies stack from Data Storage, Data Processing, Data Visualisation to Data Analytics
- Acquire skills to manage and analyse big data
- Implement key predictive modelling Algorithms on RapidMiner
- Perform Exploratory data analysis and data pre-processing techniques
- Identify the right tool for solving real life big data problems
- Get hands-on experience in using Big Data Technologies Hadoop, HBase, Hive, RapidMiner

EXAM PREPARATION

The objective of the certification examination is to evaluate the knowledge + skills acquired by the participants during the course on Big Data. The weightage in key topics of the course as follows:

- Introduction to Business Analytics [10%]
- Data/Information Architecture for Business Analytics [15]
- Data Mining Tool [5]
- Data Mining Techniques [10]
- Introduction to Big Data [10%]
- Introduction to Hadoop [15]
- Hadoop HDFS & MapReduce [15%]
- Apache HBase [10%]
- Apache Hive [10%]

Tools/Software used

- Hadoop
- HBase
- Hive
- Rapidminer